DOCUMENT RESUME

ED 088 927 TM 003 487

AUTHOR Anderson, Scarvia; And Others
TITLE Priorities and Directions for Research and Development Related to Measurement of Young Children: Report on Task 2.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Office of Child Development (DHEW), Washington, D.C.
REPORT NO ETS-PR-72-22
PUB DATE Oct 72
NOTE 37p.

EDRS PRICE MF-$0.75 HC-$1.85
DESCRIPTORS Action Research; *Children; Cultural Factors; Environmental Influences; Kindergarten Children; *Measurement; Measurement Techniques; Preschool Children; *Research; Research Utilization; Social Influences; *Testing; Testing Problems; Test Validity

ABSTRACT

A panel of 15 experts in child development, early childhood education and measurement met in September 1972 to assist the Office of Child Development in establishing priorities in improving tests and measurements for young children. A summary of the panel discussion is presented along with the specific recommendations made by the participants. The key issues under consideration were: (1) the special statistical and methodological problems of measuring the behavior of young children and the impact of their environments because of the limited response system of young children and the rapid changes that occur in early life; (2) the considerations of construct-based measurement, particularly the problems of population and ecological validity that are inherent in the use of measures with different cultural groups; and (3) the dependency of the advancement of measurement research and development on appropriate policy decisions, and the availability and training of manpower. (RH)

ED 088927

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to:
TM PS
In our judgment, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

NOV 2 1 1972

PR-72-22

PRIORITIES AND DIRECTIONS FOR RESEARCH AND DEVELOPMENT RELATED TO MEASUREMENT OF YOUNG CHILDREN

Report on Task 2 under
OCD Grant Number H-2993 A/H/O

October 1972

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY

ED 088927

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRO-
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIG-
INATING IT. POINTS OF VIEW OR OPIN-
IONS STATED DO NOT NECESSARILY REPRE-
SENT OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY

PERMISSION TO REPRODUCE THIS COPY-
RIGHTED MATERIAL HAS BEEN GRANTED BY
Dorothy Cuban
TO ERIC AND ORGANIZATIONS OPERATING
UNDER AGREEMENTS WITH THE NATIONAL IN-
STITUTE OF EDUCATION. FURTHER REPRO-
DUCTION OUTSIDE THE ERIC SYSTEM RE-
QUIRES PERMISSION OF THE COPYRIGHT
OWNER.

Priorities and Directions for Research and Development Related to Measurement of Young Children

Report on Task 2

under

(OCD Grant Number H-2993 A/0/0

This report presents a synthesis of the results of a panel discussion held in Princeton on September 19-21, 1972, and addressed to assisting the Office of Child Development in establishing priorities in improving the field of tests and measurements in early childhood. The panel members were:

Ernest M. Bernal, Jr.
Director, Bilingual Early Elementary Program
Southwest Educational Development Laboratory

Courtney Cazden
Professor of Education
Harvard University

Edgar G. Epps
Professor of Urban Education
University of Chicago

Susan W. Gray
Professor of Psychology
George Peabody College for Teachers

Henry J. Mark
Associate Professor of Pediatrics, Otolaryngology, Environmental Medicine
Johns Hopkins Hospital

Virginia Shipman
Senior Research Psychologist
Educational Testing Service

Irving E. Sigel
Professor of Psychology
State University of New York at Buffalo

Herman Witkin
Senior Research Psychologist
Educational Testing Service

Others attending the meeting were Esther Grash, Office of Child Development, and Scarvia Anderson, Samuel Ball, Ruth Ekstrom, Nathaniel Gertsbrenn, Ann Jungeblut, and Samuel Messick of Educational Testing Service.

The report was prepared by Scarvia Anderson, Samuel Messick, and Nathaniel Gertsbrenn.

Contents

	Page
Science is measurement.	1
Improvement of measures begins with improvement in conceptualization.	2
Conclusions about children and their environments will stand or fall on the basis of the adequacy with which major variables are assessed.	5
Assessment of environments is important to understanding the functions of individuals.	9
Action systems are needed for the effective utilization of research and measurement in educational and social applications.	12
Effective measurement development, research, administration, and utilization are dependent on availability and training of manpower.	16
The advancement of measurement research and development in early childhood depends upon appropriate policy decisions.	20
Summary of Recommendations.	24
Appendix: Specific Panel Recommendations.	29

Science Is Measurement

The basic task of measurement is the generic task of all science: the marshaling of evidence to support an inferential leap from an observed consistency in the empirical world to a construct that will explain that consistency. In psychometric parlance, this is the problem of construct validity. In judging the adequacy of measurement, there are important statistical and methodological criteria (such as validity and reliability) that must be satisfied, but these are simply part of the central requirement of a theoretical rationale.

Measurement of young children and their environments presents some special challenges to the statistical and methodological criteria because of the limited response system of the young child and the very rapid changes that occur in early life. However, these problems must not sidetrack the investigator from basic theoretical inquiry into the nature of child development and educational functioning--an inquiry in which measurement can play a central and organizing role.

This paper, therefore, is primarily concerned with considerations of construct-based measurement, particularly with the problems of population and ecological validity that are inherent in the use of measures with different cultural groups.

Improvement of Measures Begins with Improvement in Conceptualization

Measurement development pursued as part of a theoretical framework instead of on an ad hoc basis permits one to (a) evaluate the adequacy of the measurement in terms of the meaning of the construct, (b) consider individual score differences as representing more or less of the trait measured, and (c) compare and integrate results across studies in terms of common constructs.

If we eventually want to use measurement for practical purposes such as diagnosis and evaluation, we must be prepared to justify that use in terms of social consequences, and these cannot be evaluated without information about the meaning of the measure. No accumulation of sterile statistics can compensate for lack of understanding.

Multiple Measures and Multiple Domains

The meaning of a measure is interpreted or evolves from its pattern of relationships with other theoretically relevant measures (convergent validity) and its lack of relationship with theoretically unrelated measures (discriminant validity). Therefore, research and development on measurement must be multivariate in nature. This is a general principle of all measurement--physical, environmental, and sociological as well as psychological. It implies that to explain the meaning of a measure in full, it is important to examine its operation in domains other than the one from which it derives. The investigator interested in a psychological variable such as field independence, for example, would be interested in how it operates in many situations and across different cultural groups. This introduces the notions of population validity (the extent to which the meaning of a measure--or the results of an experiment--will generalize to other population groups), ecological validity (the extent to which generalization is possible to other environmental settings), and task validity (the extent to which the selected measurement task is representative of the external domain of interest or to other tasks sampled from the same domain). These are much more powerful conceptions of validity than the limited and simplistic criterion-oriented methodology characteristic of applied statistics.

The extent to which a measure has the same meaning or displays the same properties in the same or different groups under different conditions (including different times) is an important empirical question. It is particularly critical in early childhood research and evaluation because progression through developmental stages may involve qualitative reorganizations of psychological dimensions, thereby changing the meaning derived from the measures. Different processes may be involved in a task at different times--a test of number conservation given at age 5 may reflect intuitive perceptual understandings, whereas at age 7 it may reflect concrete operational thinking. This is the problem of continuity vs. discontinuity in measurement. Within this framework, then, it is very important for improvement of the field to focus not only upon changes of levels in performance over time but also upon changes in patterns of organization of these performances.

Important Domains and Constructs in Child Development

In the early childhood area, there are some key domains that require intensive examination to uncover or define salient constructs. In other cases, there are promising constructs that require further elucidation. Third, there is a need to search out constructs that cut across domains and offer the possibility of explaining interactive processes.

Some key domains that should be investigated are family processes, language development, affective development, coping strategies, learning processes (as opposed to outcomes), and adult decision processes related to the care and treatment of children. It is not that some of these domains lack measures (there are a great many measures of "Ability to Cope with Personal-Social Demands") but that they lack the kind of conceptualization or theoretical organization that makes possible adequate assessment of the quality and meaning of the measures.

Some promising constructs requiring further elucidation, especially at the early childhood level (much work may have been done on older populations which cannot be simply extrapolated downward), include dimensions of creativity, intelligence, and cognitive style. Other important constructs deserving further attention as a basis for sound measurement of young children are components of concrete and formal operational thinking derived from Piaget.

A major concern in the study of human development is to understand the integration of differentiated subsystems in the child and, in particular, the interplay of cognition and personality. Investigations into this area should take account of three possible forms of this interplay: mediation, interaction, and transaction. Processes in one domain may mediate functioning and development in another ("dependency" may mediate the development of "analytical skills" in cognition); variables in two or more domains may interact to determine function or development (cognition plus motivation may determine academic achievement, unpredictable from either variable alone); and an observed behavior may be so holistic in character that it represents a transaction among contributing variables of such a nature that they can no longer be discretely identified or disentangled (a child's interrupting a teacher to ask a question may represent an aggressive act, an act of dependency, or an act of cognitive coping--or all of these things at once, in which case a more complex abstraction is required).

Implications for Research and Development

What are the implications of this stress on theoretically based measurement for those who fund and encourage research and development efforts in the early childhood area? There are several:

1. Investigations involving multiple domains and multiple measures have a greater chance of advancing knowledge in the field than do studies of single constructs or measures, however global.

2. Since policy decisions to initiate, enlarge, or terminate programs are based on the results of particular evaluation studies, it is important that such studies include analyses of results across individuals, population groups, and situations.

3. When new measures are needed for research and evaluation efforts, preference should be given to those that have been derived as part of a theoretical framework rather than to ad hoc developments.

Conclusions about Children and Their Environments Will Stand or Fall on the Basis of the Adequacy with which Major Variables Are Assessed

Just as we cannot necessarily extrapolate constructs from one age to another or measures from one situation to another, neither can we necessarily extrapolate assessment principles or test theory appropriate for older ages to the measurement of young children. At the same time, we must not lower technical standards just because the subjects of study are young and because their functioning is less differentiated and more dependent on situational influences. It is still essential to make inferences from the performance of young children to underlying personality and cognitive processes, and this requires as firmly based and well supported evidence as any other kind of measurement inference--probably even more so because of the child's greater susceptibility to contextual variations.

Six Major Needs

In applying assessment procedures to young children, there are six major needs that deserve special attention:

1. The need for a systematic examination of currently accepted test theory and principles and commonly held assumptions to determine their applicability to the assessment of young children: A good example of why this is necessary can be seen in the whole body of prescription and practice that has grown up around the concept of guessing on multiple-choice items. The frequently used formula $S = R - \frac{1}{n-1} W$ (where S = score, R = number of items right, W = number wrong, and n = number of choices offered in an item) seems fairly sensible when applied to a population who have developed out of their experience some specific strategies for test taking ("If you can eliminate one or more choices as clearly wrong, guess; if you cannot eliminate any choice as clearly wrong, don't guess"). It is highly unlikely, however, that children of five or six would have developed such strategies. Similarly, we can expect relatively sophisticated test takers to recognize that the correct answer is equally likely to appear in any of the response positions. When faced with difficult items, however, young children without this insight are more likely to respond in terms of position biases or other types of response sets.

2. The need to develop procedures specific to the measurement and analysis of change: Traditional psychometric methods employed in test and scale construction emphasize indices of internal consistency and stability. They seek items that maximally discriminate between individuals at a given point in time. However, these may not be the items that are optimally responsive to change processes occurring as a function of development or educational treatment. It has been suggested that a new kind of psychometrics needs to be developed to handle the special problems associated with measurement of change. Special problems also arise in the analysis of change; we must go beyond the assessment of differences in level to investigate the possibilities of differences in structure that might signal changes in the meaning of measures across time periods. Methodological investigations into the measurement of change are especially vital to research in early childhood.

3. The need to eliminate all irrelevant measurement difficulty: Requiring memory or reading abilities in a test of social studies competency may be permissible for 12th graders because these kinds of ancillary skills are required at such a simple level that individual differences in them do not contribute to response variance. However, for younger children individual differences in such abilities are likely to be pronounced and would tend to contaminate any measure of their understanding of social studies. This kind of contamination has led to the fair charge that many achievement and ability tests are really reading tests in disguise. Other examples of irrelevant difficulties that may interfere with valid assessment include a response procedure that is almost as difficult for the child to understand as the problems posed by the test itself, or a time limit that is severely restrictive when the test task requires varying amounts of reflection by the respondents. Slavish adherence to "standardized" administration procedures has sometimes been more of a detriment than a contributor to test validity. The important thing with young children is to design test materials and arrange testing conditions in any way that will maximize the likelihood that the child will understand the task demands and respond along the dimensions intended by the examiner, dimensions intrinsic to the construct being investigated--in other words, to ensure that the test task becomes the child's task.-

4. The need to match response requirements both to the task at hand and to the relatively undifferentiated response system of the young child: At the infant level, of course, the problems of choosing meaningful response channels are exacerbated. (Ingenious investigators have turned to dimensions of the orientation reflex, for example, to obtain indicators of attention and information-processing abilities and consistencies in the infant.) At the same time, however, the ability of even very young children to respond in a variety of ways should be thoroughly explored and not underestimated out of hand.

5. The need to extend measurement standards not only to young age levels but also to non-test instrumentation: The use of the word "test" in the preceding discussion does not imply that investigators using other forms of measures such as questionnaires, observations, and interviews are relieved of obligations to demonstrate the adequacy of their techniques. However, these kinds of measures are not as typically supported with evidence on reliability and construct validity partly because investigators in these areas are not generally as immersed in psychometric thinking and partly because questionnaires and observations apparently capture behaviors in such a direct way that they are sometimes taken at face value.

6. The need to explore relationships between unobtrusive measures and standardized test procedures: Some general confusion surrounds the attempts to avoid the problems of irrelevant difficulty by substituting unobtrusive measures, since procedures such as observation are sometimes misclassified as unobtrusive. There are really two dimensions leading to four quadrants of classification here: reactive vs. nonreactive (in terms of the measurement task), obtrusive vs. unobtrusive (in terms of the measurement context). Observations are frequently nonreactive but obtrusive. (Indeed, whenever the investigator or observer is present on the scene, problems of obtrusiveness come to the fore; so also may problems of reactivity.) Standardized situational tasks observed through one-way screens are reactive and unobtrusive. Trace measures such as "noseprints on the glass" or "worn-down tiles" are neither reactive nor obtrusive.

What needs to be done is to recognize how task variations along these dimensions influence the kinds of inferences investigators are prone to make and what additional kinds of evidence may be required to support inferences in the different cases. For example, if a measure is blatantly obtrusive, what kinds of supplementary evidence need to be accumulated to ascertain whether the obtrusiveness seriously contaminates the meaning of the scores derived? This may be viewed as a special case of the problem of method variance contaminating trait variance.

The decision to use either natural or contrived settings often appears to be a matter of the investigator's taste, when it should depend on the proposed use of the "scores" in subsequent analyses. If the observational measures are to serve as dependent variables, they should be derived from standardized situations. If they are to serve as independent variables describing the program or treatment, they may be derived from naturalistic settings, although valuable predictive (independent variable) information can also stem from standardized situations. Confusion on this point may result in such anomalies as treating the number of questions a child asks in class as a descriptor both of the kind of educational process he is experiencing and the outcomes of the particular educational treatment.

It is important to add that in a systems view of the organism interacting with his environment, the labelling of variables as "dependent" or "independent" may not be as important as recognizing their interdependence. However, this view does not eliminate consideration for each variable of the logic of measurement and experimental control. For instance, in the example of question asking given above, there would be little hope of predicting individual consistencies in question-asking behavior from observations obtained in a naturalistic setting where children had widely varying opportunities and occasions to ask questions.

Implications for Research and Development

Investigations should be launched into the appropriateness and properties of measurement methods as well as into the nature of the constructs being measured, and these enterprises should proceed simultaneously. Most important is the need to match methods of measurement to both the characteristics of the constructs and the response capabilities of the subjects.

Assessment of Environments Is Important to Understanding the Functioning of Individuals

Psychometric science grew primarily out of attempts to measure characteristics of individuals, and the majority of measurement efforts over the years have been in this direction. Recently, however, there has been increasing recognition that human behavior is multiply determined by a variety of internal and external variables and that test responses, being behaviors in the small, are similarly multiply determined. Hence, if we are to understand sources of test variance and the constructs underlying test performance, we must give increased attention to the context of that performance and the environmental factors influencing it. More generally this concern embraces both the assessment of the immediate context in which the measurement of the individual takes place and the assessment of the broader environmental settings influencing educational and psychological development.

It is assumed that the assessment of environmental variables should follow the same principles of construct measurement outlined in the second section of this report. In other words, measurement development should be based on theory. However, most of the constructs underlying measurement of individuals are derived from psychological theory (particularly cognitive and personality theory) while those underlying measurement of environments are derived from sociological, economic, ecological, and social-psychological theories. Interpersonal relationships (including person-group and group-group interactions) are primarily the concern of social psychologists; interenvironmental relationships (recognizing overlapping environmental variables that impinge one on the other) are the concern of systems analysts and operations researchers. The interaction of people and environments is the growing concern of the newly evolving fields of environmental psychology and ecology.

General Environments

Typically, individuals and environments are measured separately and their interactions are studied through research. Investigators who use this strategy are presently more hampered by lack of adequate measures of the environment than of the person.

Some of the environmental areas of importance to child development where improved measurement is needed are dimensions of family process and socialization, educational programs, physical and spatial properties and constraints, and school and community life. At the same time, there is a need to reexamine some of our conventional demographic measures of socioeconomic status; if, indeed, we are to conduct investigations across population groups, we must demonstrate the comparability of meanings of such measures for the several groups.

In addition, some of the variables of person-environment interaction are coming to be conceptualized as constructs in their own right and this offers the intriguing possibility of measuring such variables directly. In fact, this is one of the most promising directions for future measurement research and development. It would permit us to take direct account, for example, of the possibility that each child in a classroom may actually be experiencing a different educational program and that each sibling may be living in a different home environment. Furthermore, we may have to recognize and measure certain processes that mediate between the individual and his environments, as in the study of social perception and personal space. In many cases, we may misconstrue the nature of relationships derived solely from measures of the individual and of people and environments that impinge upon him or even from direct measures of the interaction among them; we may have to measure the person's perception of these other people and environmental characteristics and interpret the interrelationships and interactions from the standpoint of his personal construction of the world.

The Assessment Environment

We are sometimes interested in the context of assessment primarily to identify possible threats to the validity of assessment results. This concern is especially pertinent to interpretation of measurement results obtained with very young children. While relatively wide variations in testing conditions and settings may have very little effect on the test performance of adults (especially if the assessment relates to their motivations or aspirations), they can drastically alter the performance of children. At least investigators must devise methods of assessing these test-condition variables. (This is an

addition, of course, to attempting to devise facilitating and positive contexts for testing.) Some of the kinds of context assessment that are important here are interpersonal (child-examiner or child-child if there is more than one child in the assessment setting), personal (including the child's response styles and feelings of adequacy in coping with the task demands), environmental (intrusive external events), temporal (how long does the assessment take?), physical (room arrangement, heat, light, and so on), and examiner-based (the examiner's characteristics and administration styles). Just as we need comparability of measurement of constructs across investigators to permit accumulation of knowledge and impact, so do we need comparability of methods of assessing the context of assessment to compare results investigators get using the same measures.

Implications for Research and Development

Priorities in the area of environmental assessment in child development include attention to direct measures of interaction and assessments of mediating processes as well as measures of common "main effect" environmental variables. In addition, it is important to document the immediate context of assessment to clarify possible influences on scores that may require qualifications of inferences and generalizations.

Action Systems Are Needed for the Effective Utilization of Research and Measurement in Educational and Social Applications

One of the perennial difficulties in dealing with educational and social problems is moving from research and development to practical applications of its ideas and materials--applications that are practical in economic, political, organizational, and humanistic terms. The most sophisticated approach to these difficulties is one that develops action systems that include essential components and takes account of all of the interests and concerns of the various parties to the enterprise. In education this means recognizing explicitly that approaches that do not meaningfully involve teachers and parents in developing the goals of a new curriculum project are unlikely to succeed.

What kind of action systems would be appropriate to carry out the ideas that result from research in early learning? An example of a complex system very much desired by those concerned with early learning disabilities would involve the following components:

1. An assessment battery, well-grounded conceptually and valid in terms of its predictive consequences, to identify children likely to experience educational problems and to diagnose specific deficiencies and proficiencies.

2. Guidelines for interpretation of battery results at a level of complexity appropriate to the phenomenon of interest: The discrete pieces of information from the assessment may be combined in various ways, depending upon the identification/diagnostic needs. In some cases, combinations of weighted scores may be sufficient; in others, the important thing would be not so much level as pattern of, and discrepancies in, performance.

3. Treatment specifications and prescriptions based on the assessment results or patterns for individual children: The determination of relevant specifications and prescriptions for appropriate programs that results from the diagnosis of deficiencies and proficiencies must itself be the result of extensive research and development efforts. This is probably the most important missing link in the child development field and should be given the highest priority.

4. Procedures for periodic monitoring of the progress of children in the programs and for evaluation of the effectiveness of the treatments: These procedures should include some of the same instrumentation used in the initial

assessment battery. In addition, they may include assessments of reactions to children's progress and to the programs by parents and other concerned groups.

5. Correction mechanisms keyed to the results of component 4 (above) to enable (a) new treatment prescriptions for children as predictions and diagnoses change and (b) modifications of treatment specifications to try to improve them: Because of the rapid developmental changes--both qualitative and quantitative--that are likely to occur, such a recycling component is vital in a system designed to serve the educational needs of young children.

Any action system should contain within it, from the very outset of its implementation, this kind of provision for periodic collection and analysis of evaluative information in order to effect its improvement. In some cases, this means identifying changes in conditions that might require program modification. In addition, if evaluative information is positive, it can be used to justify the continuation of a program or, if negative, to modify or terminate it on a rational basis. The inclusion of cost-effectiveness information in a program-evaluation model increases its utility for these purposes.

Program evaluation within a construct framework, if sufficiently systematic in design and execution, can qualify as research on educational process with the potentiality for contributing to the advancement of knowledge about child development and practice that that implies.

At a less ambitious level, it has been suggested that it would be of great service, especially to local educational planners, to have access to a kit of measures from which they might choose instruments to try out in their own action systems. The measures in the kit would be selected by experts from fields concerned with the assessment of young children and their environments. The experts would employ selection criteria related to such characteristics as construct validity, other kinds of validity, reliability, adequacy and clarity of administration directions, availability of related equipment, and comprehensiveness of the total collection. The last reflects the major point already made in this paper about the importance of multi-domain, multi-measure investigations, where domains include the psychological, physical, and sociological and the measures include tests, naturalistic observations, and questionnaires. The principle applies as much to action systems as to research studies. Any such kit effort as that described above would be useful ove time only to the

extent that provision was made for periodic updating of its contents, including elimination and addition of measures.

An important principle distilled from Seymour Sarason's book, The Culture of the School and the Problem of Change,* is that whenever an attempt is made to improve or change a social enterprise involving several interested parties, and this attempt takes into account all of the vested interests but one, the neglected party will rise up in an organized fashion to destroy the effort. Nowhere is Sarason more likely to be proved right than in as socially and politically sensitive an area as one involving the measurement of young children.

Preventing the effort from being scuttled, however, is not the only reason for involving all interested parties in such an enterprise. Actually, their contributions to the conceptualization of measurement-related problems and the selection and application of measures can make the results of those processes more meaningful. Teachers know better than any other group what educational actions they have the facilities and resources to undertake following the diagnosis of children's educational needs. Parents know better than any other group what educational aspirations they have for their children. Both teachers and parents frequently know better than anyone else what kinds of materials and situations the children are likely to respond to.

Using Standard-related Measures

A significant movement in educational measurement today is away from interpretation of test performance in relative (normative) terms to interpretation in terms of standards of acceptable or desirable performance. Leaders of this movement use various terms such as "criterion-referenced measurement," "domain-referenced measurement," and "measurement for mastery." What they are all saying is that for purposes of improving a child's performance it is more important to know where he stands relative to standards of accomplishment than to the performance of others (although the latter may provide significant signs of potentialities or possible problems in his development, and such signs are especially important at early age levels).

---

*Sarason, S. The culture of the school and the problem of change. Boston, Mass.: Allyn & Bacon, 1971.

The use of such standard-related measures is especially pertinent to action programs because it can be tied directly to educational prescription and intervention. Moreover, the emphasis upon standards highlights the necessity of confronting the value issues of what is "good" in a pluralistic society or whose standards will prevail. When a local program undertakes the development or use of such measures, the involvement of all the interested parties takes on special urgency and significance. They must be involved not just in the standard-setting process but also in the prior identification of the goals that the standards are meant to serve and the means whereby attempts will be made to reach them.

It should be emphasized that there are research as well as operational implications of the criterion-referenced thrust. Work needs to be done on the technical properties of such measures and their relationship to construct-based assessment. Furthermore, it needs to be generally recognized that the development of criterion-referenced measurement in contradistinction to norm-referenced measurement does not mean that normative information is not valuable. Indeed, it is unfortunate that the two measurement approaches are frequently viewed as a polarity, for not only can they usefully supplement each other but normative considerations, albeit usually implicitly, often underlie the choice of instrument content as well as the performance standards set for criterion-referenced tests.

Consequences of The Program and Effects on The People Involved

An area of special concern--and difficulty--in measurement programs is the effects of the measures and the accompanying decision processes on the people involved. These are, of course, inextricably intertwined because the effects of the measurement on people are frequently mediated by the decisions and actions of other people. In the case of the effects of measurement on young children, however, the "other people" have the unfair advantage of being larger and stronger!

Consider, first, the decision makers in measurement programs. Stress has already been placed on both (a) the essentiality of involving key parties in

the decisions about what constructs are to be assessed and how they are to be assessed in terms of the purposes of the program and (b) the importance of insisting that the assessment conditions and materials will not involve irrelevant difficulties, make some individuals feel anxious, threatened, or alienated, constitute invasions of privacy, and so on.

But there are some prior and overriding questions the decision groups must face, including the purposes and social consequences of the program and any assessment that is required for it. Furthermore, before the program is put into operation, they must specify the uses and limitations on uses to which the measurement results can be put. Typical education uses relevant to young children include instructional guidance of individuals or groups, evaluation of the effectiveness of an innovative program, and curriculum or program revision. Typical misuses that must be guarded against from the outset derive from misconceptions of the phenomena being measured ("intelligence" is a major case in point), exaggerated expectations about the infallibility of tests, tendencies to take seriously insignificant differences between scores, injudicious presentations of results (in forms not directly related to the needs of the teacher, counselor, or other interpreter), and desires to make data collected for one purpose serve other purposes for which it was not intended or particularly appropriate.

Tests and other measures can have both positive and negative effects on those who take them, administer them, and interpret them. The takers in early childhood projects are usually the children, and frequently their parents and teachers as well. Taking a test should not be an unpleasant experience for a child. In fact, if the measure is appropriately designed, the activity can be rewarding and even fun. Moreover, some tests [illegible] can provide an excellent opportunity for the teacher or other [illegible] to observe a child intensively and study his reactions and coping behaviors for insights this information may provide for future educational efforts. In addition, a good assessment battery can do much to promote among teachers and others consideration of the complexity of children and the broad range of skills, attitudes, social competencies, and so on that characterizes children's development and underlies their responses to educational and social stimuli. Experience with construct-based measures can enhance understanding of the constructs on which the

measures are based. Similarly, a good questionnaire can increase a mother's consciousness of factors, including values, important in her and her child's life.

There are numerous such examples of possible positive effects of measures on those involved with them, and the list of possible negative effects includes anxiety, stimulation of over-competitiveness, and invasion of privacy. The point is that there is a serious need for continuous consideration of potential social and personal consequences in any proposed use of measurement. These ethical issues must be squarely faced as an integral part of decision making in measurement research and application.

Implications for Research and Development

In this section dealing with action programs involving measurement of young children, the major principle is that processes of decision making about uses of measurement should occur within a rational framework that includes attention to:

1. The interdependencies of the components of the action system.
2. The priorities and goals of all of the parties to the enterprise.
3. Provision for evaluative information for the improvement and adaptation of the system.
4. Possible measurement side effects (negative and positive).
5. The decision processes themselves.
6. The ethical basis for the assessment (and the system) in terms of personal and social consequences.

Effective Measurement Development, Research, Administration, and Utilization Are Dependent on Availability and Training of Manpower

The current OCD interest in the establishment of a new profession of Child Development Associate (CDA) is a recognition of the shortage of trained personpower to assist in Head Start, Parent-Child centers, and other programs dedicated to serving young children in the United States. In terms of the focus of this report, this shortage is felt especially in the area of administering assessment instruments. It is recommended, therefore, that special programs be developed (in relationship to the CDA effort or otherwise) to train people in the skills, sensitivity, patience, flexibility, and humor that good administrators of measures for young children must have.

This is not an easy recommendation to implement. Wisdom and economics are on the side of using testers from the same communities as the children being tested, which implies a nationwide training effort. It is difficult and time-consuming to train people in assessment skills that may have to be applied to a variety of situations and instruments, and include skills in administering measures to parents, teachers, and other adults in children's lives as well as the children themselves. Perhaps, after initial training, periodic refresher courses or short-term courses to train in new measures would be required. Furthermore, not many people in any one community can expect to make testing a full-time occupation. Therefore, it is important that people be trained in other skills as well that will make them useful in a wide range of child services.

Manpower training programs also need to be developed in the instrument-development process or "art," as it is sometimes described. As we have suggested elsewhere, there is a far-from-perfect correlation between knowing what to assess and knowing how to assess it. Development of instrumentation for young children presents unusual problems that standard university tests and measurements courses do not usually cope with.

The various applications of measurement in relationship to child development require different mixes of expertise and experience. The researcher, evaluator, administrator, diagnostician, and teacher all represent specialized roles, and, while many individuals frequently are able to play several of them,

we must recognize that it is also possible and sometimes quite efficient to have different measurement-related tasks handled by different individuals trained in the specific mix of skills required. This is not to imply that their training in assessment should be separate from the other aspects of their professional training. Rather, it might be better to embed assessment in their total curriculum. However, investigations should be made into the best methods of increasing assessment-related skills and knowledge through existing or new structures.

The Advancement of Measurement Research and Development in Early Childhood Depends upon Appropriate Policy Decisions

The responsibilities for establishing supportive policies and atmosphere are shared by public agencies, private agencies, and individual professionals committed to research, development, and evaluation in the early childhood field. The agencies are asked to become well enough acquainted with the field of measurement in early childhood--and such recommendations as those included in this report--to appreciate the need to support committed researchers over time. "Committed" is related to the central importance of having measures firmly grounded in constructs and theories. "Time" refers both to the time this kind of effort takes and to the time necessary to allow children's developmental sequences to occur and be observed and studied. At the extreme, it can be mentioned that some of our best-known and respected psychological measures represent a product of all or a large part of the careers of prominent investigators.

Of course, it is frequently important to know whether a program is working and it may be impossible to wait for several years to find out. Even in such urgent situations, however, it is essential to provide enough time and support for sound instrument development/selection and the necessary accompanying rational processes. Otherwise, the report of the investigation, however prompt, can lead to wrong interpretations and unsound policy decisions.

On the other hand, individual investigators must not undertake sponsored research and evaluation studies for which time and resources are inadequate. And, when they can document their positions, it is important for them to be able to count on moral support from their institutions and professional organizations. This implies, of course, that they have been active in educating their institutions and developing organizational positions about the requirements for sensible research and evaluation efforts. At the same time, however, investigators must also come to appreciate that in a time of pressing social problems and rapid social change they no longer have the autonomy of time that some of them previously enjoyed. The point is that a workable balance must be struck, but the major problems at this stage appear to derive more from thoughtless action than from actionless thought.

Some of these problems can be avoided if agencies adopt a policy of supporting "targeted" research in an area, as opposed to directed investigations. The RFP route generally seems more applicable to the world of defense contracts than to the world of social science research.

Government and private agencies have generally not been inclined to support measurement development itself. Rather they have supported research and evaluation efforts that have included some instrument development. Viewed from the perspective of the first major recommendation in this report--for measurement development pursued as part of a theoretical framework--this general strategy is probably quite wise. However, in some instances, moving a promising instrument or measurement technique from the research setting into the field demands specific adaptation and development efforts. And it may very well be the case that the researchers who initiated the instrumentation or technique are not the best people to ready it for operational use. In such instances, it would be appropriate to support the further development work in its own right, at the same time according special respect to those whose talents lie in the direction of eliminating irrelevant difficulty, adjusting stimulus and response requirements to the subject and the purpose of the measurement, and in other ways ensuring that the measurement task becomes the task of the individual who confronts it.

The atmosphere for measurement research and development can be strongly influenced by two kinds of procedural routines--those having to do with review and with dissemination. In the first case, an agency's responsibility is to ensure that reviews are professionally sound and that the purposes of the process are fully explained to the researchers and evaluators whose work is being reviewed. Such reviews should have a formative and constructive component. If project reviews appear to serve only censorship or manipulative functions, they may have the direct consequence of limiting the direction of the investigation for political (or nonresearch) reasons or the indirect consequence of fostering so much advance concern that they lead to undue self-censorship to avoid possible difficulties. These statements are not to be construed as an indictment of external reviews of projects, procedures, and instruments; reviews are important and necessary to scientific inquiry. However, attention must be paid to making review procedures positive rather than punitive to keep from endangering the very investigations they are designed to serve.

Science thrives on public disclosure of its results. Any policies which seem to prevent or delay publication of the reports of investigations undertaken under agency auspices are viewed with alarm by most investigators. Many properly refuse to undertake projects when the results are intended for the sponsor's eyes alone. However, a more frequent problem in the area of dissemination has to do with making reports and products of sponsored investigations widely available. Commercial publishers and distributors--if they are good--are generally considered more able to ensure national publicity and dissemination than private or government organizations. However, ways must be found to overcome some of the copyright, "public domain," and royalty problems that have inhibited their performance of services in recent years. And working relationships have to be developed among reputable commercial organizations, investigators, and sponsoring agencies to stimulate the dissemination of promising products. The possibility of such avenues can greatly relieve the frustrations of researchers who have often in the past felt that some of their best ideas and inventions were sentenced to gathering dust on a shelf.

Finally, in the policy domain, we need to emphasize that the involvement of all of the parties with vested interests in the enterprise is just as important in research and development efforts as it is in action programs. This means that if research and development in early childhood is to focus on a particular minority group, every attempt should be made to involve researchers who thoroughly understand the problems of that group. Such involvement could range from minority-group direction of a project to collaboration to consultation, depending upon circumstances of time and available expertise. A sponsoring agency's obligations in this area include special efforts to let contracts for minority-group research to minority-group organizations and active encouragement of collaboration between minority researchers and other research organizations. Minority-group organizations have a concomitant obligation to keep informed about likely sources of support for investigations of special interest to them, to propose appropriate research and development efforts, and to be willing to offer their collaborative and advisory services to other research and development groups.

Implications for Research and Development Strategies

This section calls for conscious attention to the possible influences on the atmosphere for research and development of policy decisions in such areas as time and resources for investigations, amount and kind of external direction, types of projects to be supported, review and dissemination procedures, and involvement of relevant groups.

The policy-making process, it should be emphasized, has two distinct consequences: one the intended regulative effect and the other a change in the evaluative context or atmosphere of the regulated domain. This change in atmosphere affects the way people look at things, the details they select for emphasis, the interpretations they favor, and it thereby helps to determine the values of the future.

Summary of Recommendations

1. If measurement is to serve a positive purpose in the study of young children and their environments, its use must be justified in terms of social consequences and these cannot be evaluated without understanding the meaning of what is being measured. This understanding is possible only if measurement development is carried out within a theoretical framework.

The central task in all measurement, as in all science, is one of gathering evidence to support a theoretical explanation of phenomena observed in the empirical world. In psychometrics, this is the task of construct validity. Inherent in the construct validity approach to measurement is the notion that a variable cannot be measured in isolation. To find the meaning of a measure, one must examine the ways in which it relates and does not relate to other relevant measures. Thus, any investigation of one measure must involve others. Moreover, one must investigate how that measure functions in different situations and across different cultural groups.

For these reasons, it is important that investigations of childhood measures involve multiple measures and multiple domains. Further, since policy decisions to initiate or terminate programs are based largely on the results of evaluation studies, it is important that such studies include analyses of results across individuals, population groups, and situations. Finally, when new measures are needed for research and evaluation, preference should be given to those that have been developed as part of a theoretical framework rather than to ad hoc developments.

2. Current methods of measurement that have been found to be appropriate for older age groups cannot necessarily be applied to the assessment of young children. Most test-taking strategies that have become part of the older student's mental repertoire are unknown to the child of five or six. For the young child, many achievement tasks that are designed to measure competence in specific subject areas are contaminated by reading and memory requirements.

Further problems are posed by response procedures, which are often as difficult as the measurement tasks themselves, and time limits that are severely restrictive.

In addition to studying the nature of the constructs being measured, studies should be conducted to investigate currently accepted methods of measurement to determine ways of designing test materials and arranging testing conditions to ensure that the test tasks become the child's tasks. At the same time new kinds of measuring techniques may have to be developed to capture the complex behaviors of young children over time.

5. Children's responses to measurement tasks are influenced by many different factors both external and internal. If we are to understand children's responses to these tasks and the constructs being measured, we must give increased attention both to the context in which the measurement of the individual takes place and the larger environmental factors that influence the child's development.

Some specific areas of importance that should be investigated include dimensions of family process and socialization, educational programs, physical and spatial properties and constraints, and school and community life. In addition, there are a number of other factors having to do with the relationship between the individual and his environment that are important in their own right and that we may soon be able to measure directly.

Meanwhile, investigators must devise methods of assessing those aspects of testing conditions and settings that contribute to variations in assessment results among young children.

6. Effective action systems are required to make it possible to apply the ideas of research and development to practical needs in the field. A model system might include such components as these:

- An assessment battery, well-grounded in theory and valid in its prediction implications, to identify problems and specific deficiencies and proficiencies.
- Guidelines for interpreting results of the test battery.

. Specific treatments and prescriptions based on assessment results.

. Periodic monitoring of children's progress and effectiveness of the treatments.

. Procedures that would permit new prescriptions or modifications of existing ones if predictions and diagnoses change.

The success of all such programs will depend on the extent to which they reflect the priorities and goals of all those who are involved in their creation and use.

Those who assume the responsibility of translating the results of research and development into action programs such as those described above must also assume the responsibility for the social consequences of such programs in the communities they serve.

5. There is a clear need today to establish programs to train people in the personal and technical skills that are necessary in the administration of measures for young children. Such training should cover a wide range of child services and include provision for periodic refresher courses. Programs are also needed to train people in the development and application of instruments in child development enterprises.

6. If they are to create the proper climate for the advancement of measurement research in the field of early childhood, public and private agencies should become well enough acquainted with the field to support committed researchers for as long as they need to create measures based on carefully thought out constructs and theories.

Many problems of time and money can be avoided if agencies adopt a policy of supporting targeted research in an area instead of attempting to direct investigations.

Although government and private agencies have in the past been inclined to support research and evaluation efforts that included measurement development rather than the development of measurement itself, it may be necessary to support such efforts in order to move promising techniques from the research laboratory to the field.

Appropriate policy decisions pertaining to review of investigations and dissemination of products and reports are also important to creating an appropriate atmosphere for measurement research and development.

Finally, it should be emphasized that involvement of all the parties concerned with a project is every bit as important in research and development as it is in the establishment of action programs. If research focuses on a particular minority group, every attempt should be made to involve researchers who understand the problems of that group.

Appendix: Specific Panel Recommendations

A. Construct-based measurement development and research

Measurement development pursued as part of a theoretical framework.

Systematic simultaneous assessment of individuals and environments.

Longitudinal or developmental assessment of the changing organization of capacities, not just linear accretion in them.

Identification of constructs that are common to different subject groups but may need to be measured with different content and methodology.

Research to relate cognitive styles to functioning in the educational situation.

Assessment of ability to utilize skills, not just possession of them.

Instruments related to the child's ability to organize the environment --cognitive and affective; e.g., sense of competence, confidence in ability to cope, ability to tolerate failure, ability to apply alternative coping strategies, learning how and when to learn, internal locus of control.

Instruments in such universally important social-emotional areas as empathic abilities; tolerances of differences in appearance, thinking, etc.; feelings of competence; willingness to initiate actions.

Measures of representational ability in order to be able to deal with hindsight and anticipation (a mediating facility, as Piaget might say).

Good measures of children's communication processes.

Continued pursuit of differential assessment of different aspects of language development.

Development of early detection tools (school skills, minimum CNS) sensitive and specific to dysfunctions and specific learning disabilities for two critical ages, 2 1/2 and 4 1/2.

8. Technical characteristics and adequacy of measurement

Development of measurement standards particularly appropriate to assessment of young children and their environments--standards for instrument developers and users.

Research into the methodology of assessment of young children, with emphasis on variations in assumptions and theories as a function of subject age and culture.

Examination of the ecological validity of measures before extrapolation of program recommendations to other settings and groups.

Standardized situational "tests" to supplement information obtained from more conventional tests.

Disentanglement of the uses of observational measurement of independent and dependent variables; the former can be obtained in naturalistic or standardized settings, but the latter requires standardized settings and tasks.

Assessment of children's ability in ways and settings that engage realistic processes--especially vital in assessment of functions at a concrete as opposed to a formal level.

Assessment of cognitive skills through non-reading modalities.

Investigation of possible cultural bias or boundedness in construct definition as well as in measurement.

Analysis and reporting of possible "order" effects attributable to the arrangement of instruments in a battery.

Recognition of the "richness" of information that may be obtained from a measure--not just conventional scores but other potentially important data such as response sets, distractibility, etc.

Development of adequate "practice" materials for tests designed for administration to young children.

Investigations into the usefulness of both "limit" and standard testing procedures in the same setting; discrepancies between a child's performance under the two conditions may have important clinical and educational implications.

Provision for validation of constructs across settings--research, remedial, clinical, etc.

B. Technical characteristics and adequacy of measurement (cont.)

Routine investigations of administrator-variance when tests are moved from one setting (e.g., research) to another (e.g., educational program)--is the test author the only one who can get certain results?

Development of a taxonomy of valid and reliable responses that can be obtained from children from 0 to 9 to measurement tasks.

C. Conceptualization and measurement related to children's environments

Environmental measures, both specification of properties for measurement of specific environmental variables and instrumentation for universal dimensions that cut across specific environments (e.g., those that have to do with time coerciveness).

Measures capable of describing dynamic as opposed to static processes in the child's interactions with his environment.

Measures of children's experience in context (their "individual" educational programs).

Improvement of instruments used to gather demographic data (e.g., SES) and determination of comparability of meaning across population groups.

D. Action systems involving measurement

Development of strategies for the simultaneous selection of measurement variables and identification of program needs and for establishing research, development, and evaluation priorities; one strategy might involve emphasis on the overlap between research and consumer priorities and comparisons of treatment effects for different populations.

Provision for taking account of consumer needs and values in conceptualization of measurement-related problems and in the development, selection, and application of measures; consumers include those directly responsible for the welfare of the children.

Consideration of prescription as a necessary sequent to evaluation, understanding, and development of a range of alternative program options (to challenge the consumer to rational choice).

Focus in assessment interpretations on individual differences and intra-pattern analyses, as opposed to group differences and comparisons.

Investigation of obtaining diagnosis and prediction information from a single set of assessments.

Observational procedures suitable for monitoring the installation and implementation of an educational innovation.

Assessment of ability to utilize skills, not just possession of them in terms of abilities in vocal, pantomime, recorded (reading and writing), and mathematical/scientific languages predicted from auditory/vocal and visual/fine-motor skills; determination of the relationships between such discrepancies and social, emotional, and cognitive problems of children.

Tests that describe capabilities and limitations for which some "treatment" can be prescribed (e.g., criterion-referenced tests), as opposed to tests interpretable only in normative terms.

Selection, by experts, of a multi-measure, multi-domain "kit" or collection of measures from which instruments can be selected for tryout at local levels.

Survey of the actual educational decision-making processes that attend the assessment of young children, for possible insights into improving them.

Inclusion of a search for possible side-effects (positive and negative) of measures on young children.

Investigation of problems associated with "labelling" as a consequence of administration of certain child instruments.

Recognition of and capitalization on the positive side-effects on teachers of participating in instrument selection, administration, and interpretation.

D. Action systems involving measurement (cont.)

"Job analyses" of typical school learning tasks as an aid to program and instrument development/selection.

Development of self-correcting, uniform (computer-compatible) decision trees which display the decision process in the selection of teaching strategies matched to ability profiles of groups and individuals.

Study of the effects of overt and hidden cognitive skills and handicaps (and patterns thereof) on the child's scholastic achievement, social adjustment, emotional adjustment, and the family's satisfaction with school performance and the child's performance; study of the specific teaching strategies that are effective with children of different skill-handicap patterns.

Estimation of potential cost benefits of self-correcting diagnosis-treatment evaluation systems related to dysfunction and specific learning abilities at early education levels, especially in comparison with the costs of such current programs and practices as "non-promotion," Right to Read Programs, Drop-Out Prevention Programs, and Special Remedial Programs.

E. Manpower development and training

Training procedures and systems for testers and other gatherers of data about young children and their environments.

Development of subprofessional manpower to serve dual roles in individualized data acquisition, translation, and feedback processes and to act as "teacher assistants."

Improved and specific training in development of instruments for young children.

F. Research and development policy and atmosphere

Support of committed researchers over time--time sufficient to deal with the complexity of construct assessment and for developmental sequences to occur.

"Targeted" but not directed stimulation of research and development in early childhood assessment and research.

Special agency efforts to let research and development contracts to minority groups and/or to encourage collaboration between minority research groups and other research organizations.

Agency appreciation of the time it takes for rational processes and instrument development in research and evaluation efforts in early childhood; concomitant professional acceptance of the responsibility not to undertake government-sponsored research and evaluation without adequate time and resources.

Development of specific research and development priorities related to measurement of infants.

Modification of agency policy (if necessary) to allow for direct support of instrument development, especially the application of measurement expertise to promising conceptually-based research instruments.